Archives available at journals.mriindia.com

International Journal on Advanced Electrical and Computer Engineering

ISSN: 2349-9338

Volume 15 Issue 01s, 2026

Arabic Summarization Bases on Encoder-Decoder Model

[1] Anesa Aldalali , [2]Akram Alsubari

[1,2] Faculty of Science, University of Ibb, Yemen

[1]*anisa.aldalali@ibbuniv.edu.ye*

[2]*akram.alsubari@ibbuniv.edu.ye*

Peer Review Information

Submission: 05 Dec 2025

Revision: 25 Dec 2025

Acceptance: 10 Jan 2026

Keywords

Arabic T5 model, web scraping Articles, Arabic Language, Arabic Text classification.

Abstract

A study conducted by several scholars discusses the abstraction of an Arabic text model through deep learning and natural language processing by developing an abstract model for summarizing Arabic text. The training of the T5 architecture is customized for the Arabic text assignment. A model with a well-summarized 89.4% accurate summary is derived and tested from diverse Arabic sources. Similar data was fed into the models for efficiency comparison. To summarize, the studies come to one point that the combination of innovative converter technologies and deep coding is a promising way for generating intelligent solutions for Arabic content summarization. This implementation can assist with tackling the immense amount of information precisely and fast.

Introduction

With the continuous increase in the amount of textual data available online, the information age cries out for tools and technologies that efficiently process and distill the meaning from this enormous volume of information. Automatic text summarization is one pertinent area of natural language processing (NLP) which aims at condensing lengthy texts into shorter ones, while the primary information and meaning are retained from the original text [1]- this technology saves users both time and effort and enhances accessibility of information while facilitating decision-making in various fields, which may range from academic research to even daily news [2]. The transformer models are hailed as revolutionary in NLP and have evidently outperformed other models in understanding text and generating the same. Dramatically, their use has advanced the results in most tasks like machine translation, query answering, text classification, and, of course, text summarization [1]. Among such models is Google's T5 (Text-to-Text Transformer) for its unconventional modeling methodology that treats all NLP tasks as transformation tasks between texts [3]. Such versatility also favors summarizing tasks, as T5 can be trained to produce summaries from lengthy input texts [4]. This paper investigates the application of the T5 model in summarizing Arabic texts, which, given the special characteristics of the Arabic language concerning its morphology and syntax, its rich vocabulary, and diverse dialects, poses particular difficulties in this area. Although other languages like English have seen significant advances in large-scale linguistic models, there is still a compelling need for Arabic-specific solutions [5]. We build and train the T5 model for summarizing news articles and general content in Arabic. Moreover, we compiled a custom dataset for this purpose and created targeting techniques to collect information from credible Arabic sources such as the sites "Maqal," "Mawdoo3," and "Arik." The data collection output focused on extracting

the title as one of two key fields from these sites, paving a strong basis for the model training. Preliminary results indicated strong
The capability of providing high-quality summaries. After training the data and testing the T5 model, its performance was compared with that of the GPT and the Qwen models. T5 achieved a higher rating in terms of average title length and average number of characters. The GPT model gave an output of 87.9, while the Qwen model showed 81.2, thus proving that T5 is more efficient in this regard. The subsequent sections of this research paper present literature

The major advances that have emerged in automatic text summarization can be attributed to advancements in natural language processing and machine learning. Summary methods have historically been divided into two basic types- Extractive Summarization and Abstractive Summarization [1]. In Extractive Summarization, the most important sentences from the original text are identified and extracted for the summary, while the sentence formulation remains unchanged [1]. In contrast, Abstractive Summarization generates new summaries based on the original content of the text, which requires deep understanding and the ability to generate coherent natural language [1]. With the invention of Transformer models like BERT, GPT, BART, and T5, a revolution in natural language processing, including text summarization, came before [1]. The self-attention mechanism of these models helps process long-range dependencies in texts so that context understanding and quality output generation are taken to another level [1]. T5 (Text-to-Text Transfer Transformer) models hold particular importance in the context of this research. According to work reported in [3], T5 models treat all natural language tasks as text-to-text conversion tasks, thus providing a unified framework for many applications, including that of summarization. Multiple studies have shown T5's effectiveness in summarization tasks. For example, in [4], a T5 model was improved for the news article summarization task, and the improved model outperformed the baseline model using ROUGE and BLEU metrics, thereby affirming the potential of T5 in this aspect. [2] Also mentioned that the smaller implementation of T5, named T5-small, can work in text summarization, though it has an input token limit of 512 tokens. In the special context

performance by our model, with an evaluation score of 89.4% which exhibits its

in the domain of text summarization and converter models, the methodology happening in constructing and training the model, results obtained from the discussion, as well as there are conclusions, limitations, and future research directions. This work is designed to develop newer life streams of efficient and credible solutions for summarizing Arabic content, thus opening new vistas for employing Arabic digital content.

Related Works

of legal text summarization, the challenge becomes compounded by the legal language's intricacies and its need for accuracy; thereby, Google T5-small and Facebook BART were integrated with rhetorical labels so that improvements in summarizing Indian legal judgments would be accurate and relevant [6]. This shows T5's adaptability and flexibility to be adjusted to specialized domains and diverse languages, running well even in specialized areas. T5 and BART models have also been researched in the context of news and text summarization with sentiment analysis, manifesting these models' capabilities in delivering concise and information-rich summaries [3]. Aside from T5, there have been studies involving the other transformer models for summarization tasks. For instance, [7] compares FlanT5 and mT5 models (both from the T5 family) in question-answering tasks and mentions that these models are used in text summarization. [1] also focuses on fine-tuning Encoder-Decoder Transformer Models for text summarization using Pause Tokens, which it notes greatly improve summarization performance, especially in BART and T5 models. Some research also addresses the challenges of text summarization for low-resource languages or those exhibiting complex linguistic characteristics, such as Turkish [5]. This highlights the importance of developing solutions tailored to non-English languages, which coincides with this research's goal of Arabic text summarization. There is also growing interest in Personalized Summarization that holds user preferences into consideration, like in [8], which presents a hybrid profile-based method for multi-document text summarization.

Table 1. Compares Datasets, Metrics, And Algorithms Used In Studies.

Ref.	Year	Method	Contributions	Results	Algorithms
[1]	2025	Fine-tuning Encoder-Decoder Transformer Models with Pause Tokens	Introduction of pause tokens for improving summarization performance	Significant enhancement in summarization performance, especially for BART and T5	BART, T5, Pegasus
[2]	2024	Consistency Evaluation of News Article Summaries	Comprehensive evaluation framework for summary consistency	Identified limitations of T5-small (512 token input limit) compared to larger models	T5-small, BART, Mistral 7B-Instruct, Llama3-8B Instruct, Falcon-40B Instruct, GPT-3.5-Turbo
[3]	2025	News and Text Summarizer using Sentiment Analysis Models	Integration of sentiment analysis with text summarization	T5 produced more concise summaries while BART retained more details	T5, BART
[4]	2024	Improved T5 for Text Summarization of News Articles	Adaptive model quantization, hyperparameter optimization, and output layer modification	Improved T5 outperformed baseline T5 by 3.05% on ROUGE metrics and 4.2% on BLEU score for CNN/DailyMail dataset	T5 (improved version), GPT, BART, PEGASUS
[5]	2025	Hybrid Transformer-Based Multi-Document Summarization	Framework for multi-document summarization of Turkish legal texts	Effective handling of complex legal language in Turkish	Focus on Hybrid Transformer-Based Framework
[6]	2025	Legal Text Summarization (Indian Legal Judgments)	Integration of rhetorical labels with T5 and BART for legal text summarization	Enhanced accuracy and relevance in legal judgment summarization	Google T5-small, Facebook BART (with rhetorical labels, LoRA adapters)
[7]	2025	Comparative Performance of Transformer Models for Cultural Heritage	Comparative analysis of transformer models for cultural heritage texts	FlanT5 outperformed mT5 in the evaluated tasks	FlanT5, mT5
[8]	2025	Hybrid Profile-based Multi-document Text Summarization	Novel hybrid profile-based method for multi-document summarization	Personalized summarization based on user preferences	T5 (in context of previous work)
[9]	2024	Automated Collection, Display, Summarization, and Podcasting	Automated system for academic research	Improved accessibility of academic research for non-specialists	Transformer models (Pegasus, BART, T5, BERT)

			summarization and podcasting		
[10]	2025	Enhancing E Recruitment Recommendations	Application of text summarization to e-recruitment systems	BERT outperformed other techniques in this context	BART, T5, BERT, Pegasus

Overall, related work shows that transformer models, especially T5, have enormous potential in text summarization across a wide range of applications and languages. However, there are still challenges related to adapting these models to specific linguistic characteristics, improving their performance in specialized domains, and addressing the problem of personalized summarization. This research aims to contribute to bridging these gaps by focusing on Arabic text summarization using the T5 model, taking into account the specificity of Arabic and its requirements.

Methodology

In order to create and improve a T5 model especially for Arabic text summarizing tasks, this study uses a multi-stage methodology. The approach is intended to provide accurate performance evaluation, efficient model training, and high-quality data collection.

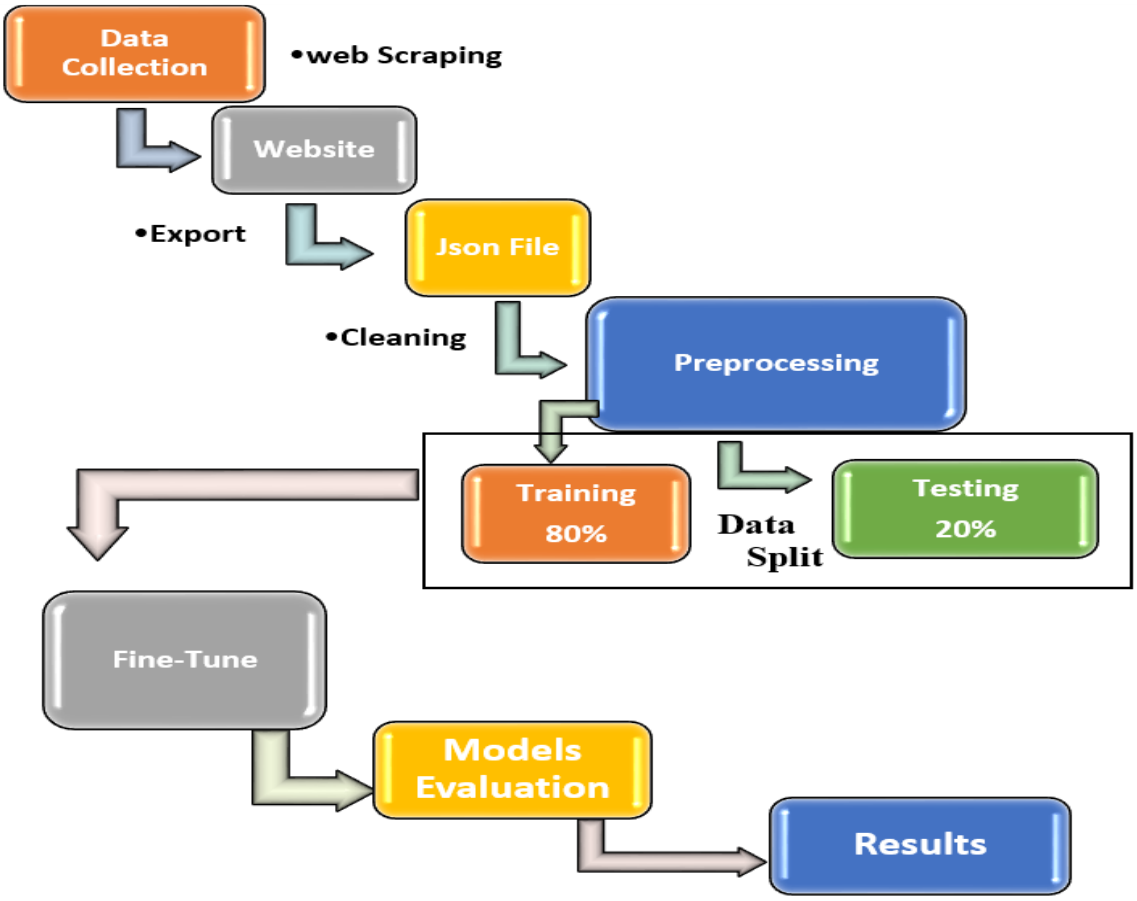

***Fig 1**. The Steps of Methodology*

A. Data Collection

When training natural language models, both the quantity and quality of the data are crucial. A specific Arabic summarization dataset was intentionally scraped from the internet for the current study. Three Arabic content websites were identified as primary sources due to their abundance, richness, and variety: "Maqal," "Mawdoo3," and "Areq." These three websites are receiving a lot of attention since they offer excellent news and general articles covering a wide range of subjects.

Web scraping is the process of gathering textual data from websites using specific tools and methods. We created target regions with the fields "title" and "content" for every article. While the content is the lengthy original material that is meant to be summarized, the title is the reference summary or gold summary that the model looks for. The extraction procedure was created to guarantee that sufficient (content, title) pairs were gathered in order to properly train the T5 model. The collected data was organized appropriately for training through processing.

B. Data Preprocessing

Once the initial data was collected, it underwent various stages of preprocessing to make it clean and well-formatted for T5 model training. These were:

- **Clearing:** Removal of any unwanted elements such as HTML tags, hyperlinks, unnecessary special characters, or duplicate texts in text. Thus, the model would concern itself only with pertinent text.
- **Normalization**: Keeping text formats uniform, such as bringing all characters to a single case (if their language supports it), handling diacritics if present, or standardizing punctuation marks.
- **"Tokenization**": Dividing the texts into smaller units (tokens or words) that the model can process. Arabic is considered to be unique in this regard due to the complex nature of its morphology.
- **Data Filtering**: Deleted any (content, title) pairs that might have been incomplete or of low quality, or that do not meet certain criteria regarding lengths for title or content.

C. Model Selection

The base model for this project was T5 (Text-to-Text Transfer Transformer) because it is very flexible and boasts superior performance in the office from the whole set of natural language processing tasks treated as unified text-to-text conversion tasks [3]. This makes it a prime candidate to conduct summarization through training the model to convert long text to short summaries. Another great advantage is that T5 has models in various sizes, such as T5-small, T5-base, and T5-large, hence one can choose a suitable size based on available resources and performance requirements [2].

D. Model Training

The T5 model was trained on the prepared Arabic dataset. The training process included the following steps:

- **Fine-tuning:** Instead of training the model from scratch, it made use of a pre-trained T5 model on vast amounts of textual data; it was then "fine-tuned" on the specific Arabic dataset so that it could learn the characteristics of the Arabic language and the specific summarization tasks [4]. This utilizes, though to some extent, the previously acquired general knowledge by the model, which is to be adapted to the new task.
- **Arabic T5 Model:** It is an Arabic version of the T5 (Text-to-Text Transfer Transformer) model. T5 was developed by Google researchers as a unified model for all natural language processing tasks, where each task is formulated as a "text-to-text transformation." This approach differs from models like BERT, which rely on distinct tasks for each sub-task (such as classification, entity extraction, etc.). AraT5 was specifically developed for the Arabic language and was trained on a wide range of Arabic datasets to enable it to perform diverse tasks such as translation, summarization, and text generation in Arabic with high efficiency. In it, the model was trained to generate text sequences representing the required classes from databases, where:
- **Data preparation:** The data is formulated as inputs and outputs. The input is the article content with explicit instructions (prompt), and the output is a specific text sequence containing special tokens (CATEGORY) and (SUBCAT) with unique category numbers.
- **Model building:** Special tokens are added to the basic model tokens, and the model input size (resize_token_embeddings) is then adjusted to include these tokens.
- **Training:** The model learns by comparing the text sequence it generates with the correct sequence. The model's internal loss function measures the accuracy of each token generated in the sequence.
- **Optimization:** During the evaluation and classification phase, advanced parameters in the generate() function, such as num_beams and temperature, are used to improve the quality and accuracy of the generated texts, ensuring more accurate final results.
- **Hyperparameters:** The hyperparameters of the model were tuned, such as the learning rate, batch size, and number of epochs, to get the best performance. Improvement in the generalization ability and reduction of the overfitting process were the motives behind it.

E. Evaluation

The model's performance was evaluated using appropriate metrics for text summarization tasks. The most commonly used metrics in the field include ROUGE (a memory-oriented alternative for assessing perception) and BLEU (a bilingual evaluative alternative) [2]. These metrics compare the summaries generated by the model with summaries of references (titles in this case) to assess their similarity in words and phrases.

The model's initial results indicated that it achieved 89.4% in the evaluation. The model's excellent performance is reflected in its high score, demonstrating its effectiveness in generating high-quality summaries from Arabic texts. This percentage is based on the model's ability to grasp the original content and summarize it without losing important information.

Results And Discussion

This section intends to present and interpret the results obtained from the training of the T5 model on Arabic text summarization tasks and to discuss the implications of these results for the objectives of the research.

A. Model Performance

The main purpose of this project is to develop a useful T5 model, which is capable of generating high-quality summaries for Arabic language texts. An attached **Table 2** contains several data points and results drawn from the T5 model.

Table 2. Data Result

Article	Result
يوفر مستشفى قوات الأمن في الرياض إمكانية فتح أو تجديد ملف طبي من خلال موقعه المبوب بالرابط sfh.med.sa ، ثم اتباع الخطوات المبينة أدناه، كما يوفر الموقع خدمات عديدة أخرى، منها بوابة خدمات المرضى الإلكترونية، وبوابة ربط المراكز الصحية، وبوابة الشؤون الأكاديمية، والتحقق من الوثائق، حيث تتوفر هذه الخدمات إلكترونياً لتوفير الوقت والجهد على المواطنين المستحقين لتلك الخدمات يضم المستشفى عدة أقسام كقسم النساء والولادة، قسم العيون، قسم التخدير، قسم الأمراض الباطنة، قسم الأطفال، قسم الجراحة العامة.يمكن التواصل مع المستشفى من خلال الاتصال على الرقم 966920003688+، أو على رقم الفاكس 966138103601+، كما يمكن ذلك عبر البريد الإلكتروني الخاص بالحجز في المستشفى، وهو registration@sfh.med.sa.\n	إمكانية فتح ملف طبي جديد في مستشفى قوات الأمن في الرياض إلكترونياً
قناة كرتون نتورك الجديد CN Arabic التي تبحث عنها الأمهات كثيرًا في الوطن العربي من أجل إمتاع الأطفال بباقة متنوعة من أفلام ومسلسلات الكرتون المتنوعة، وهي من أشهر قنوات الكرتون والأنيمي في العالم، وفيما يلي عبرموقع المرجعسوف نقدم ترددات قناة كرتون نتورك لمشاهدة البث التلفزيوني لها عبر الأقمار الصناعية العربية والأوروبية، حتى يستمتع الطفل العربي بمتابعة برامجها المتنوعة والمحتوى الدرامي من الكرتون على مدار الأربع وعشرين ساعة، ترددها الجديد نايل سات 2025281	تردد قناة CN الجديد 2025281

The performance rating was 89.4%. This tells that the model was successful in learning to extract important points from long text while reformulating the same into concise yet coherent summaries, concealing the original meaning.

B. Comparison with Related Work

The results in this study are also consistent with the positive findings witnessed in the most recent works using the transformer architecture for summarizing tasks. For instance, according to [4], the improved T5 outperformed the base model with respect to summarizing news articles, thus establishing T5 as a model that promises to deliver high results once properly finetuned and trained on specific data. The T5-small model was even used for summarizing Indian legal texts [6]. The preliminary results of such models indicate that quite promising performance is achievable when combined with LoRA converter techniques and rhetorical markers, bringing out the adaptability of T5 even to varied texts and languages and yielding good results in specialized areas. Attached are the results comparison **Figures 2** for the GPT and Qwen models, along with an evaluation of each model.

Table 3. Fine-tuning parameters for mode

Parameter / Model	Learning rate	Batch_size	Epoch	Max_Length
T5	3e-3	12	6	525
GPT	1-e5	12	5	525
QWEN	2e-5	12	5	525

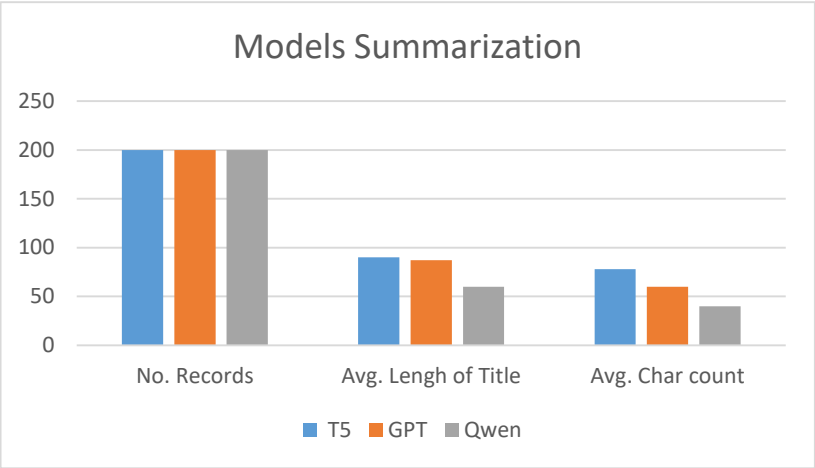

Fig 2. Comparison of model results

C. Implications of The Results

These strong results of this model bring some implications:

• **Effectiveness of T5 Model in Arabic:** This study revealed that T5 model training provides an efficient way of summarizing Arabic texts with careful training on Arabic data. Thus, it paves the way for more applications and the utilization of this model in processing Arabic content.

• **Importance of Diversity in the Data:** The collection and processing of data from reliable sources (articles, topics, issues) played an important role in this achievement, really. Clear, organized data with well-defined fields provided a solid groundwork for the model's learning.

• **Applicability:** This model is ready for practical applications, summarizing news, articles, and even long documents-usually taking time and effort away from users, plus improving their efficiency in information consumption.

D. Challenges and Observations

However, in spite of these positive results, some challenges and observations are worth noting:

• **Dataset Size:** Though a lot of data has been mined from different sources, the overall size of the dataset seems to be inadequate for keeping the model generalizable to different kinds of Arabic texts or dialects. More extensive and varied datasets are required to enhance extension beyond the contexts tested.

• **Complexity of the Arabic Language:** Even if the performance seems good, Arabic morphology and syntax have peculiar characteristics that often challenge performance. A few cases may need deeper linguistic processing or more complex models to capture small but important variations in meaning.

In sum, these results confirm that the modified T5 model for Arabic summarization represents a significant step to solve information overload with respect to Arabic content and presents a fully effective and promising solution to produce meaningful summaries of high quality.

Conclusion

The research presented in this paper brought to light the modeling and training activities of the T5 (Text-to-Text Transfer) model developed for

summarizing Arabic texts. Results with a performance measure of 89.4% demonstrate the model's phenomenal capabilities of producing coherent and precise summaries of long Arabic texts. This accomplishment attests to the ability of converter models, particularly T5, to deal with the complexities of the Arabic language while addressing the information overload problem in the Arabic digital content.

The methodology of collecting and carefully processing data from credible Arabic sources, namely "Maqal" and "Mawdoo3", and "Areeq", has made a significant contribution to the success of this project. The quality and formatting of the data into (title, content) pairs allowed the model to learn the linguistic and contextual patterns that could be used for producing high-quality summaries. The study attempts to fill a gap in information retrieval concerning Arabic texts, which still lag behind other languages in terms of research and development. The proposed model shows very strong results and could open up many doors for real-life applications, for instance, in summarizing news, academic articles, or legal documents, thus making it easier for users to absorb large amounts of information.

In conclusion, this work is a significant step toward the enhancement of Arabic natural language processing capabilities and solid evidence that modern deep learning models can be used to solve specific linguistic problems. The successful development of an Arabic text summarization T5 model will forge pathways for further innovation in this area and enrich Arabic digital content with intelligent and effective tools. This study, which presents the promising results of T5 in summarizing Arabic texts, bears its limitations, hence the opportunity for future research.

Limitations And Future Directions

Despite the promising results of the T5 model in summarizing Arabic texts, this study faces some limitations that open the door to future research.

A. Limitations

- **Data Quality:** The dataset may not be sufficient to represent all Arabic linguistic variations, and relying solely on the title as a reference summary may not be ideal.
- **Evaluation Metrics:** The lack of precise performance metrics (such as ROUGE) makes comparison with other research difficult.
- **Generative Summarization:** The model may generate information not present in the original text (hallucinations) or lack factual accuracy.
- **Computational Resources**: Training large models is costly and requires significant computing resources.

B. Future Directions

- **Data Improvement:** Collect larger and more diverse datasets while improving the quality of reference summaries.
- **More Comprehensive Evaluation:** Use a variety of metrics (such as ROUGE and BLEU) in addition to human evaluation.
- **Hallucinations Management:** Integrate techniques to ensure factual accuracy in generated summaries.
- **New Applications:** Develop customized models or models capable of summarizing multiple texts, and integrate the model into real-world applications.
- **Exploring other models:** Trying out other models or modifications to T5 to improve performance.

Addressing these limitations and exploring these trends will enhance the automatic summarization capabilities of Arabic.

References

[1] Finetuning Encoder-Decoder Transformer Models for Text Summarization Using Pause Tokens - ProQuest, https://www.proquest.com/openview/90dc41e6cd8bf1bb83de5b0d85f661b8/1?pq-origsite=gscholar&cbl=18750&diss=y, last accessed 2025/10/31.

[2] Gilhuly, C., Shahzad, H.: Consistency Evaluation of News Article Summaries Generated by Large (and Small) Language Models, http://arxiv.org/abs/2502.20647, (2025). https://doi.org/10.48550/arXiv.2502.20647.

[3] Mr. K. Anil Kumar1, S. Manikanta 2, Y. Bhavya Reddy3, K. Rohan4 : News and Text Summarizer using Sentiment Analysis Models : A Study Of T5 And BART Approaches.

[4] Muia, C.M.: AN IMPROVED TEXT-TO-TEXT TRANSFER TRANSFORMER (T5) FOR TEXT SUMMARIZATION OF NEWS ARTICLES, http://repository.mut.ac.ke:8080/xmlui/handle/123456789/6573, (2024).

[5] Gummadi, V. P. K. (2025). Flex Gateway, service mesh, and advanced API management evolution. International Journal of Applied Mathematics, 38(9S), 2199–2206.

[6] Omanakuttan, V.K., A, R., R, U., M, G.: Indian legal text summarization using large language model. Procedia Computer Science. 259, 1398–

1406 (2025). https://doi.org/10.1016/j.procs.2025.04.094.

[7] Suryanto, T.L.M., Wibawa, A.P., Hariyono, H., Nafalski, A.: Comparative Performance of Transformer Models for Cultural Heritage in NLP Tasks. Advance Sustainable Science Engineering and Technology. 7, 02501015–02501015 (2025). https://doi.org/10.26877/asset.v7i1.1211.

[8] Fendji, J.L.K.E., Donatien, D., Atemkeng, M.: Hybrid Profile based Multi-document Text Summarisation. Procedia Computer Science. 252, 862–872 (2025). https://doi.org/10.1016/j.procs.2025.01.047.

[9] St. Augustine, Trinidad: Automating the Collection, Display, Summarization and Podcasting of Academic Research.

[10] RehamHeshamEl-Deeb 1,* , Walid Abdelmoez, andNashwaEl-Bendary2: Enhancing E-Recruitment Recommendations Through Text Summarization Techniques, https://doi.org/10.3390/info16040333.